Received: 20 January 2016 | Revised: 11 May 2016 | Accepted: 23 June 2016

DOI: 10.1002/mpr.1532

WILEY

ORIGINAL ARTICLE

Cross-cultural measurement invariance of the General Health Questionnaire-12 in a German and a Colombian population sample

Matthias Romppel[1] | Andreas Hinz[2] | Carolyn Finck[3] | Jeremy Young[3†] | Elmar Brähler[2,4] | Heide Glaesmer[2]

[1] Institute for Public Health and Nursing Research, Department of Prevention and Health Promotion, Bremen University, Bremen, Germany

[2] Department of Medical Psychology and Medical Sociology, University of Leipzig, Leipzig, Germany

[3] Department of Psychology, Universidad de los Andes, Bogotá, Colombia

[4] University Medical Center, Clinic for Psychosomatic Medicine and Psychotherapy, Johannes Gutenberg University Mainz, Mainz, Germany

Correspondence
Heide Glaesmer, Department of Medical Psychology and Medical Sociology, University of Leipzig, Philipp-Rosenthal-Str. 55, 04103 Leipzig, Germany.
Email: heide.glaesmer@medizin.uni-leipzig.de

Abstract

While the General Health Questionnaire, 12-item version (GHQ-12) has been widely used in cross-cultural comparisons, rigorous tests of the measurement equivalence of different language versions are still lacking. Thus, our study aims at investigating configural, metric and scalar invariance across the German and the Spanish version of the GHQ-12 in two population samples. The GHQ-12 was applied in two large-scale population-based samples in Germany (N = 1,977) and Colombia (N = 1,500). To investigate measurement equivalence, confirmatory factor analyses were conducted in both samples. In the German sample mean GHQ-12 total scores were higher than in the Colombian sample. A one-factor model including response bias on the negatively worded items showed superior fit in the German and the Colombian sample; thus both versions of the GHQ-12 showed configural invariance. Factor loadings and intercepts were not equal across both samples; thus GHQ-12 showed no metric and scalar invariance. As both versions of the GHQ-12 did not show measurement equivalence, it is not recommendable to compare both measures and to conclude that mental distress is higher in the German sample, although we do not know if the differences are attributable to measurement problems or represent a real difference in mental distress. The study underlines the importance of measurement equivalence in cross-cultural comparisons.

KEYWORDS

configural, German, GHQ-12, measurement equivalence, metric, scalar, Spanish

1 | INTRODUCTION

1.1 | General Health Questionnaire and its psychometric properties

The General Health Questionnaire (GHQ) was first developed in 1972. Since then it has been widely used as a screening instrument for minor psychiatric morbidity, has been translated into many languages and extensively validated in different populations (Goldberg & Williams, 1988; Werneke, Goldberg, Yalcin, & Ustun, 2000). Currently the 12-item version (GHQ-12) has become the most popular version of the scale because of its brevity. The GHQ-12 is composed of six positively phrased items and six negatively phrased items. There are different scoring methods for the GHQ-12, the 4-point-Likert scale (0123), the standard scoring (0011) and the corrected scoring method (0011 for positive items and 0111 for negative items) (Rey, Abad, Barrada, Garrido, & Ponsoda, 2014). Psychometric properties and especially dimensionality of the GHQ-12 are still under debate. Although the GHQ-12 was originally designed as a unidimensional measure, several one-, two- or three-factor solutions have been found across different studies (Werneke *et al*., 2000; Rey et al., 2014; Gureje, 1991; Kalliath, O'Driscoll, & Brough, 2004; Picardi, Abeni, & Pasquini, 2001; Politi, Piccinelli, & Wilkinson, 1994; Schmitz, Kruse, & Tress, 2001; Toyabe et al., 2007; Vanheule & Bogaerts, 2005; Hankins, 2008a; Romppel, Braehler, Roth, & Glaesmer, 2013). The impact of methodological aspects on the factorial structure of the GHQ-12 has been discussed in recent years (Rey et al., 2014; Hankins, 2008a; Romppel et al., 2013). Different scoring methods substantially affect the model estimation (Rey et al., 2014; Campbell & Knowles, 2007). Moreover, there is some evidence for the impact of positively and negatively

[†]Present address: Pontificia Universidad Javeriana Facultad de Ciencias Económicas y Administrativas, Bogota, Colombia.

worded items on factorial structure (Hankins, 2008b; Ye, 2009). The first two response categories of the 4-point-Likert scaled negatively worded items ("not at all" and "not more than usual") were supposed to be ambiguous and seemed to generate some confusion for the respondents. In contrast, the positively worded items have had different response categories which seemed to be more appropriate to distinguish frequency of symptoms. Hence, scoring method and wording seem to intertwine and possibly influence the factorial structure (Rey et al., 2014; Hankins, 2008a, 2008b).

Also, since the dimensionality and psychometric properties of the GHQ-12 are still under debate, different factor solutions from the literature have been tested for the German version of the GHQ-12 in a large-scale representative sample of the German general population (Romppel et al., 2013). The confirmatory factor analyses (CFAs) revealed best fit for the one-factor model including response bias on the negatively worded items, according to Hankins (2008a). This finding further supports the importance of methodological aspects for the dimensionality of the GHQ-12. Additionally, the correlations of the two- and three-factor models with external criteria did not substantially differ, and thus these models lack substantial additional information. The superior one-factor model showed good psychometric properties (e.g. α = 0.89, item-total correlations in the upper range, response probabilities in the medium range). Regarding the associations of the unidimensional scale with several external criteria (e.g. Beck Depression Inventory [BDI], Patient Health Questionnaire [PHQ-2], 36-item short form health survey [SF-36]), the German version of the GHQ-12 seems to be a useful screening tool for the assessment of mental distress with a main focus on depressive symptoms (Romppel et al., 2013).

The Spanish version of the GHQ-12 has received considerable attention since the 1980s, and validation processes have been reported with the general population in Spain (Rey et al., 2014; Gabriel Molina, Rodrigo, Losilla, & Vives, 2014; Serrano-Aguilar et al., 2009; Sánchez-López & Dresch, 2008), and with specific population subgroups such as inpatients, women (Aguado et al., 2012), adolescents (Padron, Galan, Durban, Gandarillas, & Rodriguez-Artalejo, 2012; Lopez-Castedo & Fernandez, 2005) or users of medical facilities in Peru (Gelaye et al., 2015), Chile (Araya, Wynn, & Lewis, 1992), Colombia (Viniegras & Victoria, 1999) and Cuba (Villa, Zuluaga Arboleda, & Restrepo Roldan, 2013), among other countries. To our knowledge, there is no published effort regarding the validation of this scale with the general population in Colombia. Authors generally agree on the efficiency of the Spanish version of the GHQ-12 for the assessment of general mental health (Viniegras & Victoria, 1999): they report good internal consistency values (e.g. α = 0.89 in the Cuban study) but also raise questions regarding the dimensionality of the scale.

1.2 | Cross-cultural measurement invariance

Culture affects people in a variety of psychological domains. The impact of culture on symptom reporting in the context of mental disorders plays an important role in cross-cultural research (Dere et al., 2015; Glaesmer, Braehler, & von Lersner, 2012). In the comparison of symptoms of mental disorders or other psychological variables across different cultures or ethnic groups, we have to ensure that we compare the same construct across the different groups. Using a psychometric scale in different groups is based on the assumption that the scale measures the same construct in every group (functional equivalence). There is a general issue of measurement invariance, meaning the equivalence of a measured construct in different groups across cultures. The debate about cross-cultural research has long been focused on the importance of functional equivalence or the comparability of validity coefficients or optimum cutoff scores but the development of measurement invariance tests and advanced statistical tools instigated more rigorous tests of measurement invariance (Chen, 2008). Nowadays the most frequently used technique for testing measurement invariance is multiple-group CFA (Chen, 2008).

There are three aspects of measurement invariance that can be tested using CFAs (Dere et al., 2015; Chen, 2008). (1) Configural invariance, as the most basic level of invariance, means that similar but not identical factors are measured in the different groups. In this sense the same items have to be associated with the same factors in each group, but factor loadings can differ across groups. If configural invariance is not met, the assessment instrument does not assess the same construct across the different groups. (2) Metric invariance means that the factor loadings of the different items are identical in the different groups, and the unit of measurement is identical. This level of equivalence is required to make meaningful comparisons of predictive relationships across groups. (3) Scalar invariance tests whether an item has the same point of origin (intercept) across different groups. This level is a precondition for the comparison of group means (Chen, 2008).

The GHQ-12 has been applied in cross-cultural comparisons. For instance, the World Health Organization (WHO) study of mental illness in general health care applied the GHQ-12 as a screening tool in 15 centers with 10 different language versions around the world. Among these 15 centers were two German centers (Mainz, Berlin) and one center in a Spanish speaking country (Santiago de Chile). The optimum thresholds for case definition varied substantially across different centers (Goldberg et al., 1997). Moreover, validity coefficients (sensitivity, specificity, positive predictive value [PPV], receiver operating characteristics [ROC]) were different across centers (Goldberg et al., 1997). Since the cutoff scores to achieve optimum sensitivity and specificity differ considerably (Goldberg, Oldehinkel, & Ormel, 1998), the application of stratum-specific likelihood ratios were recommended instead of fixed thresholds (Furukawa, Goldberg, Rabe-Hesketh, & Ustun, 2001; Furukawa & Goldberg, 1999). In recent years innovative CFA-based techniques had been implemented to test measurement invariance of psychometric instruments. The majority of cross-cultural studies do not check measurement equivalence of assessment instruments. From a rigorous psychometric perspective, the results of such cross-cultural studies are not reliable (Glaesmer et al., 2012; Chen, 2008). Although there are some cross-cultural psychometric analyses for the GHQ-12 (Furukawa, Goldberg, Rabe-Hesketh, & Ustun, 2001; Furukawa & Goldberg, 1999; Goldberg et al., 1997; Goldberg, Oldehinkel, & Ormel, 1998), CFA-based analyses on measurement invariance of the GHQ-12 are lacking to date: Thus, our study has two aims:

1. To investigate configural invariance by testing the factorial structure of the GHQ-12 in a Colombian population sample and to compare it with the findings from a German population sample (Romppel et al., 2013).

2. To investigate metric and scalar invariance across the German and the Spanish version of the GHQ-12 in two large-scale population samples.

2 | METHODS

2.1 | Subjects

2.1.1 | Colombian population sample

This sample consisted of adult people (18 years and above) of the general population of Colombia. The research market company "Brandstrat Inc." conducted the interviews in the eight main cities of Colombia: Barranquilla, Bogota, Bucaramanga, Cali, Cartagena, Manizales, Medellin, and Pereira. Each Colombian city is divided into barrios (quarters) with different mean socio-economic status (SES) of the inhabitants (SES ranging from 1 = very low to 6 = very high) which are characterized by the mean socio-economic level of the inhabitants. The sampling procedure adopted in this survey assured that each stratum (with corresponding barrios) was representatively included in the sample. Within each barrio, the participants were randomly selected. In case of non-response, another eligible participant from the same stratum was asked. This technique yielded a stratum distribution in the study sample identical with that of the general population. Due to this procedure, the resulting sample can be assumed to be representative of the urban population of Colombia living in private houses. Trained interviewers performed the survey. They asked eligible participants to take part in the study, and in case of affirmation they gave them a booklet with several questions and questionnaires and asked them to fill them in. After finishing, the interviewers reviewed the booklet for missing data and asked the people to complete the questionnaires in case of missing data (except household income).

A total of 2,372 people were contacted; ultimately, 1,500 people responded (the fieldwork ended in March 2012) with complete data sets (response rate of 63%). The interviewers did not obtain any information in case of non-participation. Therefore, we have no data on reasons of non-participation. The Ethics Committee at the Universidad de los Andes approved the study, and informed consent was obtained from all participants.

2.1.2 | German representative population sample

A representative sample of the German general population was selected with the assistance of a demographic consulting company. The area of Germany was separated into 201 sample areas representing the different regions of the country. Households of the respective area and members of this household fulfilling the inclusion criteria (age at or above 14, able to read and understand the German language) were selected randomly by Kish-selection-grid technique. The Kish-selection-grid-technique is aimed to sample individuals on the doorstep among household residents. The system is devised so that all individuals in a household have an equal chance of selection. The sample is representative in terms of age, gender, and education. A first attempt was made for 3,194 addresses, of which 3,108 were valid. If not at home, a maximum of four attempts were made to contact the selected person. Furthermore, 872 subjects (28.1%) refused participation, 137 subjects (4.4%) were not reached after four attempts, and 10 subjects (0.3%) refused participation because of severe health problems. All subjects were visited by a study assistant, informed about the investigation, and self-rating questionnaires were presented. The assistant waited until participants answered all questionnaires and offered help if the meaning of questions was not clear. A total of 2,066 people between the ages of 14 and 93 years agreed to participate, completing the self-rating questionnaires in November and December 2002 (participation rate: 66.5%). Of these, 25 subjects were excluded from the following analyses because of incomplete data, and another 64 participants under the age of 18 years were excluded with respect to comparability of the samples. A dataset of 1,977 people is included in this study. Table 1 gives an overview of the demographic characteristics of both samples.

2.2 | Instruments

The GHQ-12 (Romppel et al., 2013; Schmitz, Kruse, & Tress, 1999) with a 4-point Likert-scale (0123) was applied in a German and in a Colombian population sample. The total score of the GHQ-12 ranges from 0 to 36, with higher scores representing higher levels of mental distress.

A psychometric analysis of the German version has already been published. The CFAs revealed a superior one-factor model with good psychometric properties (e.g. internal consistency: $\alpha = 0.89$; item-total correlations in the upper range; response probabilities in the medium range) (Romppel et al., 2013).

The Spanish version used corresponds to the original translation (Muñoz, Vásquez and Rodriguez, 1979, cited in Viniegras & Victoria, 1999). Sánchez-López and Dresch (2008) reported an α value of 0.76 for their entire sample of 1,001 adults from the general population in Spain. As reported earlier, authors in Latin America agree on the psychometric qualities of the scale: $\alpha = 0.84$ for a sample of patients in Colombia (Villa et al., 2013) and $\alpha = 0.89$ for a sample of working adults in Cuba (Viniegras & Victoria, 1999). In the current study internal consistency is $\alpha = 0.83$.

2.3 | Statistical analysis

To test the configural invariance for the factorial structure of the GHQ-12 in the Colombian sample, we mimicked the analytical approach followed in a recent study in a German general population sample (Romppel et al., 2013). Thus, we modeled CFAs according to five different models that have been proposed, using maximum likelihood estimation with robust standard errors in Mplus 6.1. Model 1 represents the three-dimensional conception of the GHQ-12, with three latent variables (social dysfunction, anxiety/depression, and loss of confidence) and six, four and two measured variables loading onto them. Since a latent variable that is represented using only two indicators is locally under-identified, an equality constraint on the two loadings associated with the latent variable can be placed, following the recommendation of Little, Lindenberger, and Nesselroade (1999). Model 2 depicts the two-factor model, with two latent variables (social dysfunction and anxiety/depression) and six measured variables loading onto each. Model 3 also represents a three-dimensional conception, but with

TABLE 1 Characteristics of the German and Colombian sample[a]

		German sample[b]				Colombian sample		
		Total (*N* = 1,977)	Male (*N* = 925)	Female (*N* = 1,052)		Total (*N* = 1,500)	Male (*N* = 724)	Female (*N* = 776)
Age	*M*	49.8	48.3	51.2	*M*	41.8	42.0	41.7
	SD	17.4	16.7	17.8	SD	16.2	16.8	15.7
Age groups	18–24 years	9.0% (178)	10.8% (100)	7.4% (78)	14–24 years	18.6% (279)	20.3% (147)	17.0% (132)
	25–34 years	13.6% (269)	13.8% (128)	13.4% (141)	25–34 years	18.8% (282	17.7% (128)	19.8% (154)
	35–44 years	18.2% (360)	17.9% (166)	18.4% (194)	35–44 years	19.1% (286)	18.2% (132)	19.8% (154)
	45–54 years	16.9% (334)	17.5% (162)	16.3% (172)	45–54 years	17.8% (267)	16.3% (118)	19.2% (149)
	55–64 years	18.6% (368)	19.7% (182)	17.7% (186)	55–64 years	14.4% (216)	14.9% (108)	13.9% (108)
	65–74 years	15.3% (302)	16.3% (151)	14.4% (151)	65–74 years	8.3% (124)	9.3% (67)	7.3% (57)
	≥ 75 years	8.4% (166)	3.9 (36)	12.4% (130)	≥ 75 years	3.1% (46)	3.3% (24)	2.8% (22)
Education	No qualifications	2.1% (41)	1.2% (11)	2.9% (30)	<11 years	17.3% (259)	15.3% (111)	19.1% (148)
	Less than 10 years	45.8% (905)	44.6% (413)	46.8% (492)	11 years	36.3% (544)	32.6% (236)	39.7% (308)
	10 years	35.1% (693)	36.1% (334)	34.2% (359)	12–14 years	17.7% (266)	21.0% (152)	14.7% (114)
	More than 10 years	17.1% (338)	18.1% (167)	16.3% (171)	≥15 years	28.7% (431)	31.1% (225)	26.5% (206)
Net household income/month	< 750 €	11.2% (211)	9.8% (87)	12.4% (124)	< 400,000 COP[c]	11.5% (145)	10.1% (63)	12.9% (82)
	750 to 1249 €	28.8% (545)	25.3% (225)	31.9% (320)	400,000–799,999 COP[c]	32.0% (403)	31.6% (196)	32.5% (207)
	1250 to 1999 €	36.2% (684)	38.9% (346)	33.7% (338)	800,000–1,599,999 COP[c]	25.2% (317)	24.8% (154)	25.60% (163)
	≥ 2000 €	23.9% (452)	26.1% (232)	22.0% (220)	≥ 1,600,000 COP[c]	31.2% (393)	33.5% (208)	29.0% (185)
GHQ-12	*M* (SD)	9.70 (4.94)	9.23 (4.85)	10.11 (4.99)	*M* (SD)	7.14 (4.57)	6.96 (4.65)	7.31 (4.48)
total	Range	0–36	0–36	0–34	Range	0–29	0–29	0–27

[a]Demographic characteristics of both samples are not completely comparable because of different information assessed in both studies (e.g. years of education).

[b]Sample characteristics as published in Romppel et al. (2013) corrected for exclusion of participants under the age of 18 years.

[c]COP = Columbian Pesos (100.000 COP = about 27 €).

"cope", "stress", and "depression" as latent variables and four, three and five measured variables. Model 4 represents the one-dimensional conception of the GHQ, with all 12 items defined as indicators of a single factor. Finally, we tested the unidimensional model described by Hankins (2008a) as Model 5. In this model the GHQ-12 was modeled as a measure of one construct, but with correlated error terms on the negatively formulated items, modeling response bias. This model was therefore identical to Model 4, but it contains correlations between the error terms on the negative items. The construct reliability was calculated for each factor as the squared sum of the standardized factor loadings divided by the squared sum of the standardized factor loadings plus the sum of the variance unexplained by the factor. The construct reliability is a measure for the extent to which the indicators of a factor share common variance; values greater than 0.7 indicate good reliability. The average variance extracted (AVE) was calculated as the sum of the squared standardized factor loadings divided by the sum of the squared standardized factor loadings plus the sum of the variance unexplained by the factor. The AVE is a measure of the extent to which the variance in the indicators is accounted for by the factor; values greater than 0.5 indicate good convergent validity. The Fornell–Larcker ratio (Fornell & Larcker, 1981) is the ratio of the AVE of a factor and the squared value of the highest correlation of this factor with another factor. A ratio smaller than one indicates good discriminant validity. As fit indices of the models, the χ^2 value, the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) are reported. A good fit is indicated by values larger than 0.95 for TLI and CFI, and values smaller than 0.06 for RMSEA and smaller than 0.08 for SRMR (Hu & Bentler, 1999).

In a second step, we tested metric and scalar invariance, starting with a free-baseline model in which only the parameters of a referent item are constrained to be equal across the groups. First, we chose the referent item by running constrained-baseline versus augmented model comparisons for each item and identifying the item with the highest loading on the common factor, following the suggestion of Stark, Chernyshenko, and Drasgow (2006). Second, we compared the free-baseline model with a nested model where all item loadings are constrained to be equal across groups as a test for metric invariance. Third, as a test for scalar invariance, we compared the free-baseline model with a nested model where all item loadings and all item intercepts are constrained to be equal across groups. Finally, to test for partial metric invariance, we compared the free-baseline model with nested models where one item loading at a time (in addition to the referent item) is constrained to be equal across groups. All comparisons between baseline and nested models applied chi-square difference testing using the Satorra–Bentler scaled chi-square (Satorra & Bentler, 2001). For the tests for partial metric invariance we used a Bonferroni corrected critical p-value of 0.0045 (0.05/11; the total number of tests was 11) to reach a nominal significance level of 5%.

3 | RESULTS

Table 1 gives an overview of the characteristics of the samples. The German sample (Romppel et al., 2013) and the Colombian sample are fairly representative of the general population of these countries in terms of age and gender. The corresponding percentages of the Colombian general population can be obtained from published census data (Departamento Administrativo Nacional de Estadística, DANE, 2012). The German sample has a higher mean age, higher household income, and a higher educational level than the Colombian sample. Of the German sample 53.2% (n = 1,052) and of the Colombian sample 51.7% (n = 776) are female. In the German sample mean total scores of the GHQ-12 are higher than in the Colombian sample, and range is also higher in the German sample (0–36 versus 0–29). Both in the German and the Colombian sample women report higher total GHQ-12-scores than men (see Table 1).

3.1 | Configural invariance

The dimensionality of the German version of the GHQ-12 has already been investigated. Five different factor solutions have been tested using CFAs in the German sample. The one-factor model, including response bias on the negatively worded items according to Hankins (2008a), showed superior fit (Romppel et al., 2013). To test configural invariance of the German and the Spanish versions of the GHQ-12, the five models were then tested in the Colombian sample. Table 2 shows the results of the factor analyses in the Colombian sample. The model fit for Model 4 is inadequate. Model 3 seems unsatisfactory because of low construct reliabilities and high Fornell–Larcker ratios for two of the three factors. The factors "stress" and "depression" are highly correlated (r = 0.95) and thus are hardly discriminable. While Model 2 has a moderate fit, both Models 1 and 5 show a good fit, with some fit indices slightly favoring the one or the other model. Because Model 5 showed superior fit in the German sample and is among both models with the best fit in the Colombian sample, it seems justifiably to state configural invariance of the GHQ-12 in both samples.

3.2 | Metric and scalar invariance of the GHQ-12

Metric and scalar invariance were tested by comparing an unconstrained (free-baseline) multiple group CFA model to nested models in which (a) the item loadings (for metric invariance) and (b) the item loadings and intercepts (for scalar invariance) were constrained to be equal between the two groups. Item 5 was chosen as the referent item in the models, because it had the highest loading in a fully constrained model and an insignificant chi-square difference test when allowing the parameters to vary. In the unconstrained multiple group free-baseline model (Table 3), item 7 ("unhappy and depressed") shows the highest factor loading in the German sample, while item 11 ("enjoy day-to-day activities") shows the highest factor loading in the Colombian sample. Taken as a whole, in the Colombian sample, compared to the German sample, there are higher factor loadings for the positively worded items and lower factor loadings for the negatively worded items. The global tests for metric invariance and scalar invariance both fail with significant χ^2 difference tests ($p < 0.001$ for both tests) (Table 4). Thus, the factor loadings and intercepts are not equal between the two groups. The results of the tests for partial metric invariance (Table 4) point to the fact that the items with loadings differing between the two groups are especially among

TABLE 2 Standardized factor loadings and goodness-of-fit statistics for five alternative confirmatory factor analysis (CFA) models of the General Health Questionnaire (GHQ-12) in the Colombian sample

	Model 1			Model 2		Model 3			Model 4	Model 5
	Social dysfunction	Anxiety/ depression	Loss of confidence	Social dysfunction	Anxiety/ depression	Cope	Stress	Depression	Global	Global
1. Able to concentrate	.46			.46		.46			.38	.46
2.Capable of making decisions	.63			.63		.69			.41	.63
3. Face up to problems	.64			.63		.63			.45	.64
4. Lost sleep over worry		.58			.56		.57		.53	.30[a]
5. Constantly under strain		.71			.69		.69		.64	.35[a]
6. Could not overcome difficulties		.70			.69			.69	.66	.41[a]
7. Unhappy and depressed		.74			.73			.71	.68	.38[a]
8. Loss of self-confidence			.80		.68			.68	.65	.39[a]
9. Thinking of self as worthless			.65		.57			.57	.55	.34[a]
10. Play useful part in things	.54			.54		.56			.40	.54
11. Enjoy day-to-day activities	.68			.68			.52		.56	.68
12. Reasonably happy	.60			.60				.47	.51	.60
Construct reliability	.76	.78	.69	.76	.82	.68	.62	.76	.83	.78
Average variance extracted	.35	.47	.53	.35	.43	.35	.36	.40	.30	.24
Fornell–Larcker ratio	0.8	1.3	1.1	0.9	0.7	0.9	2.5	2.3	—	—
Correlations between factors		.53	.50		.55		.57	.54		
			.78					.95		
Fit-statistics of the model										
χ^2 *	172.2			239.3		481.5			708.9	144.3
df	51			53		51			54	39
standardized χ^2	3.4			4.5		9.4			13.1	3.7
BIC	30226.0			30329.3		30768.1			31176.0	30262.4
CFI	0.96			0.93		0.85			0.77	0.96
TLI	0.94			0.92		0.80			0.71	0.94
RMSEA (90% CI)	0.04 (0.03–0.05)			0.05 (0.04–0.06)		0.08 (0.07–0.08)			0.09 (0.08–0.10)	0.04 (0.04–0.05)
SRMR	0.04			0.04		0.08			0.08	0.04

Note: BIC, Bayesian Information Criterion; CFI, Comparative Fit Index; TLI, Tucker–Lewis Index (non-normed fit index); RMSEA, root mean square error of approximation; 90% CI, limits of the 90% confidence interval for RMSEA; SRMR, standardized root mean square residual.

[a]Error terms are allowed to covary.

*All p values <0.001.

TABLE 3 Results of multiple group confirmatory factor analysis (free-baseline model)

	Unstandardized factor loadings		Intercepts	
Item	German sample	Colombian sample	German sample	Colombian sample
1. Able to concentrate	0.794	1.087	0.986	0.848
2.Capable of making decisions	0.946	1.322	0.640	0.693
3. Face up to problems	0.717	1.331	1.008	0.754
4. Lost sleep over worry	1.099	0.917	0.727	0.744
5. Constantly under strain	1.000	1.000	0.710	0.710
6.Could not overcome difficulties	1.106	1.034	0.599	0.530
7. Unhappy and depressed	1.166	0.978	0.632	0.573
8. Loss of self-confidence	0.902	0.865	0.527	0.385
9. Thinking of self as worthless	0.853	0.632	0.442	0.250
10.Play useful part in things	0.633	1.274	0.930	0.682
11. Enjoy day-to-day activities	0.782	1.538	1.084	0.776
12. Reasonably happy	0.986	1.319	1.034	0.686

TABLE 4 Model fits of free and constrained models and tests of measurement invariance

Model	χ^2	df	standardized χ^2	BIC	CFI	TLI	RMSEA (90% CI)	$\Delta\chi^2$ [a]	Δdf	*p* Value
Free-baseline	642.6	79	8.1	63067.7	0.94	0.90	0.06 (0.06–0.07)			
Metric invariance [b]	785.7	90	8.7	63815.0	0.92	0.89	0.07 (0.06–0.07)	149.2	11	$p < .001$
Scalar invariance [c]	1304.3	101	12.9	64490.4	0.87	0.83	0.08 (0.08–0.09)	764.4	22	$p < .001$
Partial metric invariance [b]										
Item 1	649.7	80	8.1	63074.4	0.94	0.90	0.06(0.06–0.07)	6.7	1	.01
Item 2	650.2	80	8.1	63692.5	0.94	0.90	0.06(0.06–0.07)	7.5	1	.006
Item 3	669.5	80	8.4	63722.8	0.94	0.90	0.07(0.06–0.07)	31.4	1	<.001 *
Item 4	647.0	80	8.1	63684.3	0.94	0.90	0.06(0.06–0.07)	2.8	1	.095
Item 5	(reference item)									
Item 6	645.6	80	8.1	63681.6	0.94	0.90	0.06(0.06–0.07)	0.3	1	.56
Item 7	647.7	80	8.1	63684.8	0.94	0.90	0.06(0.06–0.07)	3.5	1	.06
Item 8	644.4	80	8.1	63681.2	0.94	0.90	0.06(0.06–0.07)	0.2	1	.65
Item 9	648.9	80	8.1	63688.3	0.94	0.90	0.06(0.06–0.07)	5.6	1	.02
Item 10	671.7	80	8.4	63726.0	0.94	0.90	0.07(0.06–0.07)	34.1	1	<.001 *
Item 11	680.6	80	8.5	63738.3	0.93	0.89	0.07(0.06–0.07)	53.4	1	<.001 *
Item 12	651.4	80	8.1	63690.6	0.94	0.90	0.06(0.06–0.07)	9.2	1	.002 *

Note: BIC, Bayesian Information Criterion; CFI, Comparative Fit Index; TLI, Tucker–Lewis Index (non-normed fit index); RMSEA, root mean square error of approximation; 90% CI, limits of the 90% confidence interval for RMSEA.

[a]Satorra–Bentler scaled chi-square difference test against baseline model (first row).

[b]Loadings constrained to be equal.

[c]Loadings and intercepts constrained to be equal.

*p < 0.0045 (Bonferroni corrected significance level).

the positively worded items. Specifically, the items that have higher factor loadings in the Colombian sample and lead to a significant χ^2 difference test are item 3 ("face up to problems"), item 10 ("play useful part in things"), item 11 ("enjoy day-to-day activities"), and item 12 ("reasonably happy").

4 | DISCUSSION

Our study aimed at testing configural, metric and scalar invariance of the German and the Spanish version of the GHQ-12 in two large-scale population based samples.

Five common factorial models of the GHQ-12 were previously tested using CFAs in the German population sample. In these analyses a one-factor-model including response bias on the negatively worded items showed the best fit (Romppel et al., 2013). To test configural invariance of the German and the Spanish version of the GHQ-12, we repeated this analytical approach in the Colombian sample. Therefore the five different factorial models were tested. Finally, the one-factor-model, including response bias on the negatively worded items according to Hankins (2008a), was chosen because it was one of two models with comparable good fit. Thus we can state the configural invariance of the German and the Spanish version of the GHQ-12, which implies that both versions of the GHQ-12 have a comparable

factorial structure. Comparable results concerning the factor structure of the Spanish version of the GHQ-12 have been found by some studies (Rey et al., 2014; Gabriel Molina et al., 2014; Aguado et al., 2012), while other studies have reported a multidimensional factor structure (Sánchez-López & Dresch, 2008; Padron et al., 2012; Gelaye et al., 2015).

To test metric and scalar invariance for the German and the Spanish version of the GHQ-12, multiple group CFA with item loadings (for metric invariance) and item loadings and intercepts (for scalar invariance) were constrained to be equal in both groups. Overall we found higher factor loadings for the positively worded items and lower factor loadings for the negatively worded items in the Colombian sample compared to the German sample. Factor loadings and intercepts were not equal across both groups. Thus the German and the Spanish version of the GHQ-12 show no overall metric and scalar invariance. The test for partial metric invariance points to the fact that the factor loadings differ especially for the positively worded items, and that these differences specifically contribute to the lack of metric invariance. Although both versions of the GHQ-12 show configural invariance and seem to measure the same construct, metric and scalar invariance are lacking. Thus the unit of measurement is not identical, and the point of origin is not the same. Thus from a psychometric perspective a comparison of group means has to be unjustifiable (Chen, 2008). In the Colombian sample the factor loadings for positively worded items are higher and for negatively worded items are lower than in the German sample. It seems as if the Colombians respond to a more positively connoted construct ("The glass is half full") and the Germans respond to a more negatively connoted construct ("The glass is half empty"). In this sense the different patterns in the factor loadings seem to represent a culturally caused difference. Wording effects are a well-documented phenomenon in psychometric evaluations, and the importance of positively and negatively worded items of the GHQ-12 has been investigated before (Hankins, 2008b; Ye, 2009; Wang & Lin, 2011). The wording effect means that a positive or negative formulation of items influences interpretation of these items and the response of the participants. Our study reveals configural invariance referring to the unidimensional model with response bias on the negatively worded items according to Hankins (2008a), but metric invariance is lacking with a focus on the factor loadings of the positively and negatively worded items. Therefore, wording seems not only important for the factorial structure and configural invariance but also for metric invariance. Similar problems with combining positive and negative items have been reported before (e.g. Solis Salazar, 2015; van Sonderen, Sanderman, & Coyne, 2013), that seem to outweigh the possible advantages. In addition, there exists evidence that negatively worded items are interpreted differently cross-culturally (e.g. Schmitt & Allik, 2005). This seems to plea for the avoidance of wording differences (positive versus negative) in the construction of psychometric instruments – not only in cross-cultural settings.

Even though our study is based on two large-scale population based samples, some critical points with respect to representativeness and comparability have to be mentioned. The German sample was collected similarly in rural and urban areas and is representative for age and gender for the entire adult population living in private households in Germany. In contrast, the Colombian sample was collected in eight main cities of Colombia and is representative for the urban Colombian population, but might not sufficiently represent the rural population. In Colombia most people live in urban areas, but the underrepresentation of rural areas might cause bias, and the generalizability of the results might be a matter of debate.

Nevertheless, the results clearly show that a critical analysis of measurement invariance is essential in order to avoid prematurely interpreting mean differences between groups as an indicator of different levels of mental distress.

REFERENCES

Aguado, J., Campbell, A., Ascaso, C., Navarro, P., Garcia-Esteve, L., & Luciano, J. V. (2012). Examining the factor structure and discriminant validity of the 12-Item General Health Questionnaire (GHQ-12) among Spanish postpartum women. *Assessment*, *19*(4), 517–525.

Araya, R., Wynn, R., & Lewis, G. (1992). Comparison of 2 self administered psychiatric questionnaires (GHQ-12 and SRQ-20) in primary care in Chile. *Social Psychiatry and Psychiatric Epidemiology*, *27*(4), 168–173.

Campbell, A., & Knowles, S. (2007). A confirmatory factor analysis of the GHQ12 using a large Australian sample. *European Journal of Psychological Assessment*, *23*(1), 2–8.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018.

Departamento Administrativo Nacional de Estadística (DANE). (2012). Censo 2005. http://www.dane.gov.co/index.php/poblacion-y-demografia/censos [10 April 2016].

Dere, J., Watters, C. A., Yu, S. C.-M., Bagby, R., Ryder, A. G., & Harkness, K. L. (2015). Cross-cultural examination of measurement invariance of the Beck Depression Inventory-II. *Psychological Assessment*, *27*(1), 68–81.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Furukawa, T. A., & Goldberg, D. P. (1999). Cultural invariance of likelihood ratios for the General Health Questionnaire. *Lancet*, *353*(9152), 561–562.

Gabriel Molina, J., Rodrigo, M. F., Losilla, J. M., & Vives, J. (2014). Wording effects and the factor structure of the 12-item General Health Questionnaire (GHQ-12). *Psychological Assessment*, *26*(3), 1031–1037.

Gelaye, B., Tadesse, M. G., Lohsoonthorn, V., Lertmeharit, S., Pensuksan, W. C., Sanchez, S. E., … Williams, M. A. (2015). Psychometric properties and factor structure of the General Health Questionnaire as a screening tool for anxiety and depressive symptoms in a multi-national study of young adults. *Journal of Affective Disorders*, *187*, 197–202.

Glaesmer, H., Braehler, E., & von Lersner, U. (2012). Culture-sensitive diagnostics in research and practice. State of knowledge and development potential. *Psychotherapeut*, *57*(1), 22–28.

Goldberg, D. P., & Williams, P. (1988). *A User's Guide to the General Health Questionnaire*. Basingstoke: NFER-Nelson.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine*, *27*(1), 191–197.

Goldberg, D. P., Oldehinkel, T., & Ormel, J. (1998). Why GHQ threshold varies from one place to another. *Psychological Medicine*, *28*(4), 915–921.

Furukawa, T. A, Goldberg, D. P., Rabe-Hesketh, S., & Ustun, T. B. (2001). Stratum-specific likelihood ratios of two versions of the General Health Questionnaire. *Psychological Medicine*, *31*(3). 519–529.

Gureje, O. (1991). Reliability and the factor structure of the Yoruba version of the 12-item General Health Questionnaire. *Acta Psychiatrica Scandinavica*, *84*(2), 125–129.

Hankins, M. (2008a). The factor structure of the twelve item General Health Questionnaire (GHQ-12): Results of negative phrasing? *Clinical Practice and Epidemiology in Mental Health*, *4*, 10.

Hankins, M. (2008b). The reliability of the twelve-item general health questionnaire (GHQ-12) under realistic assumptions. *BMC Public Health*, *8*, 355.

Hu, Y. J., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling*, *6*(1), 1–55.

Kalliath, T. J., O'Driscoll, M. P., & Brough, P. (2004). A confirmatory factor analysis of the General Health Questionnaire-12. *Stress and Health*, *20*(1), 11–20.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, *4*(2), 192–211.

Lopez-Castedo, A., & Fernandez, L. (2005). Psychometric properties of the Spanish version of the 12-item General Health Questionnaire in adolescents. *Perceptual and Motor Skills*, *100*(3), 676–680.

Muñoz, P. E., Vázquez, J. L., & Rodríguez, F. (1979). Adaptación española de general health questionnaire (GHQ) de Goldberg. *Archivos de neurobiologiá*, *42*, 139–158.

Padron, A., Galan, I., Durban, M., Gandarillas, A., & Rodriguez-Artalejo, F. (2012). Confirmatory factor analysis of the General Health Questionnaire (GHQ-12) in Spanish adolescents. *Quality of Life Research*, *21*(7), 1291–1298.

Picardi, A., Abeni, D., & Pasquini, P. (2001). Assessing psychological distress in patients with skin diseases: Reliability, validity and factor structure of the GHQ-12. *Journal of the European Academy of Dermatology and Venereology*, *15*(5), 410–417.

Politi, P. L., Piccinelli, M., & Wilkinson, G. (1994). Reliability, validity and factor structure of the 12-item General Health Questionnaire among young males in Italy. *Acta Psychiatrica Scandinavica*, *90*(6), 432–437.

Rey, J. J., Abad, F. J., Barrada, J. R., Garrido, L. E., & Ponsoda, V. (2014). The impact of ambiguous response categories on the factor structure of the GHQ-12. *Psychological Assessment*, *26*(3), 1021–1030.

Romppel, M., Braehler, E., Roth, M., & Glaesmer, H. (2013). What is the General Health Questionnaire-12 assessing? Dimensionality and psychometric properties of the General Health Questionnaire-12 in a large scale German population sample. *Comprehensive Psychiatry*, *54*(4), 406–413.

Sánchez-López, M. d. P., & Dresch, V. (2008). The 12-Item General Health Questionnaire (GHQ-12): Reliability, external validity and factor structure in the Spanish population. *Psicothema*, *20*(4), 839–843.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514.

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, *89*(4), 623–642.

Schmitz, N., Kruse, J., & Tress, W. (1999). Psychometric properties of the General Health Questionnaire (GHQ-12) in a German primary care sample. *Acta Psychiatrica Scandinavica*, *100*(6), 462–468.

Schmitz, N., Kruse, J., & Tress, W. (2001). Improving screening for mental disorders in the primary care setting by combining the GHQ-12 and SCL-90-R subscales. *Comprehensive Psychiatry*, *42*(2), 166–173.

Serrano-Aguilar, P., Ramallo-Farina, Y., Del Mar Trujillo-Martin, M., Raul Munoz-Navarro, S., Perestelo-Perez, L., & De Las Cuevas-Castresana, C. (2009). The relationship among Mental Health Status (GHQ-12), Health Related Quality of Life (EQ-5D) and Health-State Utilities in a general population. *Epidemiologia e Psichiatria Sociale – An International Journal for Epidemiology and Psychiatric Sciences*, *18*(3), 229–239.

Solis Salazar, M. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–199.

Stark, S., Chernyshenko, E. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306.

Toyabe, S. I., Shioiri, T., Kobayashi, K., Kuwabara, H., Koizumi, M., Endo, T., … Someya, T. (2007). Factor structure of the General Health Questionnaire (GHQ-12) in subjects who had suffered from the 2004 Niigata-Chuetsu earthquake in Japan: A community-based study. *BMC Public Health*, *7*, 715.

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS One*, *8*(7), e68967.

Vanheule, S., & Bogaerts, S. (2005). Short communication: The factorial structure of the GHQ-12. *Stress and Health*, *21*(4), 217–222.

Villa, G., Zuluaga Arboleda, C., & Restrepo Roldan, L. F. (2013). Propiedades psicometricas del Cuestionario de Salud General de Goldberg GHQ-12 en una institución hospitalaria de la ciudad de Medellín. *Avances en Psicología Latinoamericana*, *31*(3), 532–545.

Viniegras, G., & Victoria, C. R. (1999). Manual para la utilización del cuestionario de salud general de Goldberg. Adaptación Cubana. *Revista Cubana de Medicina General Integral*, *15*, 88–97.

Wang, L., & Lin, W. P. (2011). Wording effects and the dimensionality of the General Health Questionnaire (GHQ-12). *Personality and Individual Differences*, *50*(7), 1056–1061.

Werneke, U., Goldberg, D. P., Yalcin, I., & Ustun, B. T. (2000). The stability of the factor structure of the General Health Questionnaire. *Psychological Medicine*, *30*(4), 823–829.

Ye, S. Q. (2009). Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. *Personality and Individual Differences*, *46*(2), 197–201.

How to cite this article: Romppel M, Hinz A, Finck C, Young J, Brähler E, Glaesmer H. Cross-cultural measurement invariance of the General Health Questionnaire-12 in a German and a Colombian population sample. *Int J Methods Psychiatr Res.* 2017;26:e1532. https://doi.org/10.1002/mpr.1532